Appplied Data Science Capstone (week 4)

## Introduction and description oft he Problem

Berlin is considered to be one of the most popular and hippest cities both nationally and internationally. It offers many possibilities both for working and living. Last but not least, the diverse cultural offerings and the exciting nightlife attract many to the city. The cost of living is also relatively low in comparison to other metropolises around the world and makes Berlin even more attractive, so that young people are increasingly drawn to the city. Nevertheless, in recent years it has become apparent that not only has the cost of living risen in the city, but above all affordable housing is becoming increasingly scarce. As in many other metropolitan areas, rents for apartments in Berlin have risen in recent years and there is no end in sight, so that young people who come to Berlin as career starters are looking in vain for accommodation in the hip districts.

If one considers the average rent of the individual districts of the city, one also finds that the residential location is not necessarily related to the co-prices and that there must therefore be other factors that justify a district as attractive and thus the rent. Supposedly hip districts are particularly expensive while pure residential districts are pushed into the background.

This project therefore examines rental prices in Berlin and its individual districts and investigates the question: What are hip districts with expensive rents characterised by? How do neighborhoods look and what are the top venues in an expensive district as that consider the hip and fancy.

## Data and Method

To solve the problem, we will need the following data

- List of neighbourhoods in Berlin and it's neighbourhoods
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data: Sources of data is Wikipedia which contains a list of neighbourhoods in Berlin. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods
- Venue data, particularly data related to the districts:We will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developersto explore the neighbourhood regarding their top 100 venues. sing the Foursquare data in order to characterize the popular and expensive districts to determine how similar or dissimilar they are to each other and what makes them „hip".
- Resident location: good, medium, simple: These variables indicate to what extent a district within the city is attractive as a place to live and one could assume that districts with good residential locations tend to have expensive rents.
- Population density: The more inhabitants, the less space and the more expensive living space.

- Living space: The more living space, the more expensive the rent (whereby it is neglected here that the living space is related to how large a district is and over how much space is available)
- Number of residential buildings: He more residential buildings, the less the rent should be.

These assumptions will be evaluated through correlation and regression analysis so that appropriate variables can be selected to cluster the districts.

Method

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.